

Attorney Docket: YOR920000435US1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Patent Application

Applicant(s): Rigoutsos et al.
Docket No.: YOR920000435US1
Serial No.: 09/712,638
Filing Date: November 14, 2000
Group: 1631
Examiner: Cheyne D. Ly

I hereby certify that this paper is being deposited on this date with the U.S. Postal Service as first class mail addressed to the Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450

Signature: [Signature] Date: July 19, 2004

Title: Unsupervised Building and Exploitation of Composite Descriptors

TRANSMITTAL LETTER

Mail Stop Appeal Brief - Patents
Commissioner of Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

Submitted herewith are the following documents relating to the above-identified patent application:

1. Response to Notification of Non-Compliance with 37 C.F.R. §1.192(c); and
2. Corrected Appeal Brief (original and two copies).

In the event of non-payment or improper payment of a required fee, the Commissioner is authorized to charge or to credit **IBM Corporation's Deposit Account No. 50-0510** as required to correct the error. Duplicate copies of this letter and two copies of the Corrected Appeal Brief are enclosed.

Respectfully,

Michael J. Chang
Attorney for Applicant(s)
Reg. No. 46,611
Ryan, Mason & Lewis, LLP
1300 Post Road, Suite 205
Fairfield, CT 06824
(203) 255-6560

Date: July 19, 2004



IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Patent Application

5
Applicant(s) Rigoutsos et al.
Docket No.: YOR920000435US1
Serial No.: 09/712,638
Filing Date: November 14, 2000
10 Group: 1631
Examiner: C.D. Ly

I hereby certify that this paper is being deposited on this date with the U.S. Postal Service as first class mail addressed to the Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450
Signature: *[Signature]* Date: July 19, 2004

Title: Unsupervised Building and Exploitation of Composite Descriptors

15
RESPONSE TO NOTIFICATION OF NON-COMPLIANCE WITH 37 C.F.R. §1.192(c)

20 Mail Stop Appeal Brief-Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

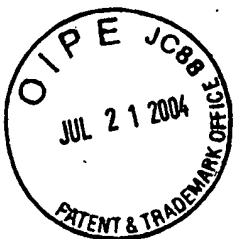
25 Sir:

In response to the Notification of Non-Compliance with 37 C.F.R. §1.192(c), dated June 22, 2004, Applicants submit herewith, in triplicate, a Corrected Appeal Brief.

Respectfully submitted,

30 Dated: July 19, 2004

[Signature]
Michael J. Chang
Attorney for Applicant(s)
Reg. No. 46,611
Ryan, Mason & Lewis, LLP
1300 Post Road, Suite 205
Fairfield, CT 06824
(203) 255-6560

AF
ITW

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Patent Application

5 Applicant(s) Rigoutsos et al.
Docket No.: YOR920000435US1
Serial No.: 09/712,638
Filing Date: November 14, 2000
10 Group: 1631
Examiner: C.D. Ly

I hereby certify that this paper is being deposited on this date with the U.S. Postal Service as first class mail addressed to the Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450

Signature: *Vina Manu* Date: July 19, 2004

Title: Unsupervised Building and Exploitation of Composite Descriptors

CORRECTED APPEAL BRIEF

Mail Stop Appeal Brief - Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

20 Sir:

Appellants hereby appeal the final rejection dated December 4, 2003, of claims 1-12, 23 and 25 of the above-identified patent application.

REAL PARTY IN INTEREST

The present application is assigned to International Business Machines Corporation, as evidenced by an assignment recorded on April 12, 2001 in the United States Patent and Trademark Office at Reel 011699, Frame 0944. The assignee, International Business Machines Corporation, is the real party in interest.

RELATED APPEALS AND INTERFERENCES

There are no known related appeals and interferences.

STATUS OF CLAIMS

Claims 1-12 and 25 stand finally rejected under 35 U.S.C. §112, second paragraph, as allegedly indefinite for failing to particularly point out and distinctly claim the subject matter of the

invention. Claims 1-8, 10-12, 23 and 25 stand finally rejected under 35 U.S.C. §102(b) as allegedly unpatentable over the article entitled “GenBank” by Benson et al. No claims have been allowed.

STATUS OF AMENDMENTS

5 An Amendment After Final Rejection, canceling claims 13-22, 24 and 26, was filed on March 3, 2004, along with a Notice of Appeal. The Examiner’s Advisory Action, issued on March 19, 2004, indicated that the amendment has been entered.

SUMMARY OF INVENTION

10 The present invention provides techniques for the unsupervised building and exploitation of composite descriptors (Specification, page 1, line 6).

The present invention provides for the determination, in an unsupervised manner, of additional members for a family that is defined initially through exemplar sequences (Specification, page 4, lines 25-27 and page 6, lines 5-7). Being unsupervised, the present techniques proceed without
15 any information related to the exemplar sequence defining the family, without aligning the exemplar sequences, without prior knowledge of any patterns in the exemplar sequences and without knowledge of the cardinality or characteristics of any features that may be present in the exemplar sequences (Specification, page 4, line 25, through page 5, line 4 and page 6, line 20, through page 7, line 9). For example, in one aspect of the invention, a method is used to take a set of unaligned sequences and
20 discover several or many patterns common to some or all of the sequences (Specification, page 5, line 5-7 and page 6, lines 13-14). These patterns can then be used to determine if candidate sequences are members of the set, e.g., family, of sequences (Specification, page 5, lines 7-8 and page 6, lines 12-15). The present methodologies may be employed as part of a system or article of manufacture (Specification, page 10, line 20, through page 12, line 3).

ISSUES PRESENTED FOR REVIEW

i. Whether claims 1-12 and 25 are properly rejected under 35 U.S.C. §112, second paragraph, as being allegedly indefinite for failing to particularly point out and distinctly claim the subject matter of the invention; and

30 ii. Whether claims 1-8, 10-12, 23 and 25 are properly rejected under 35 U.S.C. §102(b) as being allegedly unpatentable over Benson et al. (hereinafter “Benson”).

GROUPING OF CLAIMS

Claims 1, 3, 4, 23 and 25 stand or fall together. Claim 2 stands or falls alone. Claim 5 stands or falls alone. Claims 6 and 7 stand or fall together. Claim 8 stands or falls alone. Claim 10 stands or falls alone. Claim 11 stands or falls alone. Claim 12 stands or falls alone.

5

ARGUMENT35 U.S.C. §112, second paragraph, rejections

Claims 1-12 and 25 are rejected under 35 U.S.C. §112, second paragraph, as being allegedly indefinite for failing to particularly point out and distinctly claim the subject matter of the invention. Specifically, in the final Office Action at page 5, paragraph 16, the Examiner asserted that,

10

the phrase “candidate sequence” or [the] term “patterns” causes the claim[s] to be vague and indefinite. It is unclear what criteria are being used to determine that a sequence is a candidate sequence. Is it sequence identity, similarity or distribution?

15

In the Advisory Action, Continuation Sheet, paragraphs 6 and 7, the Examiner further asserted that,

[regarding] the phrase “candidate sequence”, . . . it is unclear whether said “candidate sequence” is a member of a family which requires sequence information to be known, or “the exemplar sequence wherein no information is known.” . . .

20

[Regarding] “patterns,” . . . it is unclear whether “patterns” refer to a specific sequence of symbols or information directed to the expression of said sequences that resulted in distinct patterns corresponding to said sequences. It is well known in the art that sequences of nucleic acid molecules has [sic] been widely used for discovering patterns of expression. One of ordinary skill in the art would not know whether “patterns” is directed to specific string sequences or information directed to a particular nucleic acid sequence such as expression of said sequences that resulted in distinct patterns corresponding to said sequences.

25

Appellants respectfully disagree with the Examiner’s assertions and submit that, given the present claims and supporting disclosure, one of ordinary skill in the art would understand the concepts of a candidate sequence and patterns, as well as how the two concepts interrelate with one another. Thus, one of ordinary skill in the art would be able to use these limitations to properly ascertain the metes and bounds of the present claims.

30

Regarding the phrase “candidate sequence,” by way of example only, the specification illustrates techniques for determining additional members for a family, the family being initially defined through exemplary sequences. The techniques proceed without any information related to the

35

exemplary sequences defining the family, without aligning the sequences, without prior knowledge of any patterns in the exemplary sequences and without knowledge of the cardinality or characteristics of any features that may be present in the exemplary sequences. See page 6, lines 5-11 of the specification. The additional members may be determined by analyzing candidate sequences. See page 8, lines 11 of
 5 the specification.

The statements of the Examiner (presented above) regarding whether or not knowledge about the candidate sequence is used, confuses the basic concept of a candidate sequence. Most simply put, a candidate sequence is any sequence that is a candidate for inclusion in a particular family. By way of example only, Webster's Collegiate Dictionary, tenth edition, defines "candidate" as, likely or
 10 suited to be chosen for something specified. The specification clearly presents and supports this definition.

Regarding the term "patterns," Appellants respectfully submit that the specification also clearly sets forth the concept of patterns. By way of example only, as will be described below, patterns common to some or all of the sequences (each sequence being a series of characters) in a family of
 15 sequences may be discovered. Any sequence of symbols that can be described as a linear stream of events, such as, DNA, proteins, languages and numbers can be used. See page 6, lines 13-14 and page 8, lines 15-17 of the specification.

Further, the specification also clearly sets forth how patterns are used to determine if a candidate sequence(s) is a member of a particular family. By way of example only, patterns are
 20 discovered that are common to some, or all, of the sequences in a set, e.g., a family, of sequences (which can include all members of the family). The patterns can then be used in determining whether a candidate sequence(s) is a member of the family. See, page 6, lines 12-16 and page 9, lines 3-5 of the specification.

25 Prior art rejections

Claims 1-8, 10-12, 23 and 25 are rejected under 35 U.S.C. §102(b) as allegedly unpatentable over Benson. Specifically, the Examiner asserted that Benson teaches discovering patterns. Namely, in the final Office Action beginning at page 6, paragraph 21, the Examiner submitted that,

30 Benson et al. discloses [that] the NCBI has created the UniGene collection of unique human genes. UniGene starts with human entries in the PRI division of GenBank, combines these with human ESTs and creates clusters of sequences that share virtually identical 3' untranslated regions. (citations omitted).

The Examiner then argued that the cluster of sequences “is consistent with the critical limitation of “patterns” as loosely defined by the instant specification.” See Advisory Action, Continuation Sheet, paragraph 11.

M.P.E.P. §2131 indicates that, to anticipate a claim, a reference must teach each and every element of the claim. Appellants respectfully submit that Benson does not teach each and every element of independent claims 1 (from which all pending dependent claims ultimately depend), 23 and 25. For example, Benson does not disclose or suggest (i) discovering patterns, and (ii) determining if a candidate sequence comprises a predetermined number of the patterns, as required by each claim.

Discovering patterns:

Benson does not teach discovering patterns. According to the teachings of Benson, sequences “that share virtually identical 3 untranslated regions” are clustered as a way to reduce the number of individual sequences in the database. The Examiner suggests that these clustered sequences somehow comprise patterns, as in the present claims. Assuming, arguendo, that these untranslated regions may be in some way considered patterns, there clearly is no teaching in Benson directed to a step for discovering these cluster regions, or any other regions, that might be considered to comprise patterns. Further, from the teachings presented in Benson, it appears that these untranslated regions are apparent in the sequences (e.g., basically comprising an untranslated segment in the same position (3 end) of each sequence) and are merely used to group sequences. As such, Benson does not teach discovering patterns. For these reasons alone, Benson does not disclose or suggest the teachings of independent claims 1, 23, 25 or any claims depending therefrom.

The Examiner further argued that,

“[o]ne of the most frequent uses of GenBank is sequence similarity searching.” . . . “[the] NCBI offers the BLAST family of search programs to perform fast searching with rigorous statistical methods for judging the significance of matches.” (final Office Action, page 7, paragraph 22) (citations omitted).

The Examiner also asserted in the Advisory Action, Continuation Sheet, paragraph 12, that,

ESTs (set of sequence patterns) provides the major source of new gene discoveries (candidate sequence) via BLAST searches against the dbEST (discovering patterns). Further, it is well known in the art that BLAST sequence similarity searching has been widely used for discovering similar sequences (patterns). (citations omitted).

BLAST, however, is not relevant to the teachings of the instant claims. First, BLAST does not have anything to do with the processing of sequences in a set of sequences. Namely, BLAST

does not generate anything, pattern or otherwise, from the sequences in a set of sequences. BLAST is in fact a query-driven method that involves processing a query sequence to aid in finding matches with that query in a database of sequences. Specifically, BLAST takes a query sequence and processes it to generate sequential k -tuples, wherein the value of k is fixed. The k -tuples generated are then compared to sequences in a database. Appellants, in their Amendment After Final Rejection, highlighted this fact that BLAST is a query-driven method centered around processing a query sequence. See, for example, Amendment After Final Rejection, page 8, lines 19-21. However, the subsequent Advisory Action is silent as to this point. Therefore, Appellants respectfully resubmit that the use of BLAST does not anticipate discovering patterns common to a plurality of sequences in a set of sequences and then determining if a candidate sequence comprises a number of the patterns. For this reason alone, Benson does not disclose or suggest the teachings of independent claims 1, 23, 25 or any claims depending therefrom.

Secondly, BLAST would not be a suitable program to determine whether certain patterns exist in a candidate sequence. As mentioned above, BLAST functions by generating tuples, e.g., segments, from a query sequence. Trying to match one or more patterns with a sequence, by processing the sequence (as in BLAST) would make finding any matches nearly impossible. Therefore, the teachings of Benson are inconsistent with the present teachings. For this reason alone, Benson does not disclose or suggest the teachings of independent claims 1, 23, 25 or any claims depending therefrom.

Predetermined number of patterns:

As mentioned above, Benson does not teach discovering patterns. Benson further does not teach determining if a candidate sequence comprises a predetermined number of patterns. Appellants, in their Amendment After Final Rejection, addressed this point, stating that “Benson . . . does not teach predetermining the number of patterns the candidate sequence should comprise.” See Amendment After Final Rejection, page 8, lines 23-24 (emphasis added). The subsequent Advisory Action is, however, silent as to this point. Appellants respectfully resubmit that Benson does not disclose or suggest setting a threshold parameter, such as a predetermined number of matching patterns, for sequence comparison. Benson is simply directed to managing a database of, e.g., sequences, and merely discloses general sequence comparison methods. None of these methods disclose or suggest setting such a threshold parameter. For that reason alone, Benson does not disclose or suggest the teachings of independent claims 1, 23, 25 or any claims depending therefrom.

Further, with regard to claim 2, this claim recites that the patterns common to a plurality of the set of sequences comprise test patterns and that the sequences in the set of sequences comprise

test sequences and further that the step of determining if a candidate sequence comprises a predetermined number of the patterns comprises the step of determining if there are candidate patterns in the candidate sequence that match all of the predetermined number of test patterns. Appellants respectfully submit that Benson does not disclose or suggest any of these limitations.

5 The Examiner, in the final Office Action, page 7, paragraph 22, asserted that,

 Unigene [sic] starts with human entries in the primate (PRI) division of GenBank, combines these with human ESTs and creates clusters of sequences that share virtually identical 3' untranslated regions (3' UTRs), as in claims 2 and 7. (citations omitted).

10

Appellants, however, fail to see how this teaching at all discloses or suggests test patterns, test sequences or determining if there are candidate patterns in the candidate sequence that match all of the predetermined number of test patterns. Such a teaching is not at all present in Benson.

15 With regard to claim 5, this claim recites that if the candidate sequence comprises the predetermined number of patterns, the candidate sequence is added to the set of sequences to create a new set of sequences and the step of discovering is performed on the new set of sequences. Appellants respectfully submit that Benson does not teach this limitation. The Examiner, in the final Office Action, page 7, paragraph 22, asserted that, "NCBI builds GenBank primarily from the direction [sic] submission of sequence data from authors and secondarily from scanning the journal literature, as in claim 5." (citation omitted).

20

Appellants, however, fail to see how this teaching at all discloses or suggests adding the candidate sequence having the predetermined number of patterns to the set of sequences, to create a new set of sequences and then discovering patterns in the new set of sequences. The cited section of Benson simply teaches that the GenBank database is built up of sequence data from sources, such as author submissions and journal literature.

25

With regard to claims 6 and 7, these claims recite that each sequence comprises a series of symbols and each pattern comprises a plurality of positions, some of the plurality of position each comprising at least one expected symbol and other of the plurality of positions comprising positions which may be occupied by any sequence character. The at least one expected symbol may be a plurality of expected symbols. The Examiner, in the final Office Action, page 7, paragraph 22, asserted that, "Benson et al. discloses GenBank contained [sic] 602,072,354 nucleotide bases from 920,588 different sequences, as in [claim]... 6." (citation omitted). Further, as highlighted above in conjunction with the Examiner's rejection of claim 2, the Examiner asserted that entries in the PRI division of GenBank are

30

combined with human ESTs to create clusters of sequences that share virtually identical 3' UTRs. See, for example, final Office Action, page 7, paragraph 22.

Appellants, however, fail to see how this teaching at all discloses or suggests having patterns some positions of which comprise at least one expected symbol (which may be a plurality of expected symbols) and others which comprise positions which may be occupied by any sequence character. Such teachings are not at all present in Benson.

With regard to claim 8, this claim recites that determining if each of the plurality of patterns is statistically significant further comprises selecting one of the patterns, determining if a probability that that one selected pattern occurs in a sequence meets a predetermined threshold and continuing to select additional patterns until each pattern has been selected. The Examiner, in the final Office Action, page 7, paragraph 22, asserted that,

WWW access to BLAST currently offers two interfaces, a 'Basic' version with default search parameters and an 'Advanced' option which allows customization of the parameters. A new graphical version called PowerBLAST, designed for rapid analysis and annotation of large contigs of genomic sequence data. [sic] Since the three-dimensional structure information has been linked to the set of protein sequences, users can easily determine as [sic] set of sequence neighbors for a given sequence and then locate and visualize structures for members of the neighbor set, the above disclosures anticipate the limitations of claims 8 and 10-12. (citations omitted).

Appellants, however, fail to see how this teaching at all discloses or suggests determining if each pattern is statistically significant by selecting one of the patterns and determining if the probability that the pattern occurs in a sequence meets a predetermined threshold (then repeating the process until each pattern has been selected). In fact, the cited section of Benson does not at all relate to determining the statistical significance of patterns using a predetermined threshold.

With regard to claim 10, this claim recites that determining if each of the plurality of patterns is statistically significant further comprises removing instances of each of the patterns from the set of sequences to create a new set of sequences and performing the step of discovering on the new set of sequences. As highlighted above in conjunction with the rejection of claim 8, the Examiner broadly refers to the 'Basic' and 'Advanced' versions of BLAST and to PowerBLAST. However, Appellants, fail to see how these cited teachings of Benson at all disclose or suggest determining the statistical significance of each of a plurality of patterns by creating a new set of sequences and performing the step of discovering on the new set of sequences.

With regard to claim 11, this claim recites that, if any of the patterns is statistically significant, a statistically significant pattern is selected and a composite descriptor is modified to include that selected pattern (if the selected pattern is not already part of the descriptor) and continuing to select statistically significant patterns until all statistically significant patterns have been selected. As highlighted above in conjunction with the rejection of claim 8, the Examiner broadly refers to the 'Basic' and 'Advanced' versions of BLAST and to PowerBLAST. However, Appellants, fail to see how these cited teachings of Benson at all disclose or suggest modifying a composite descriptor to include selected statistically significant patterns.

With regard to claim 12, this claim recites selecting a predetermined threshold (which indicates how many of the sequences should contain a pattern for the pattern to be considered common), discovering patterns that are common to the predetermined threshold of sequences, if there are no patterns common to the predetermined threshold of sequences, decreasing the predetermined threshold and performing (until the predetermined threshold is less than a predetermined amount) the step of discovering patterns, if any, that are common to the predetermined threshold of sequences and the step of if there are no patterns common to the predetermined threshold of sequences, decreasing the predetermined threshold. As highlighted above in conjunction with the rejection of claim 8, the Examiner broadly refers to the 'Basic' and 'Advanced' versions of BLAST and to PowerBLAST. However, Appellants, fail to see how these cited teachings of Benson at all disclose or suggest this teaching of using a predetermined threshold to discover a plurality of patterns common to a plurality of the sequences in the set of sequences.

The remaining rejected dependent claims are believed allowable for at least the reasons identified above with respect to the independent claims.

The attention of the Examiner and the Appeal Board to this matter is appreciated.

Respectfully,



Michael J. Chang
Attorney for Applicant(s)
Reg. No. 46,611
Ryan, Mason & Lewis, LLP
1300 Post Road, Suite 205
Fairfield, CT 06824
(203) 255-6560

Date: July 19, 2004

APPENDIX

1. A method comprising the steps of:
providing a set of sequences, wherein the sequences are not aligned;
5 discovering a plurality of patterns common to a plurality of the sequences; and
determining if a candidate sequence comprises a predetermined number of the patterns.
2. The method of claim 1, wherein the patterns common to a plurality of the set of
sequences comprise test patterns, wherein the sequences in set of sequences comprise test sequences,
10 and wherein the step of determining if a candidate sequence comprises a predetermined number of the
patterns comprises the step of determining if there are candidate patterns in the candidate sequence that
match all of the predetermined number of test patterns.
3. The method of claim 1, further comprising the step of determining if each of the plurality
15 of patterns is statistically significant.
4. The method of claim 1, wherein the step of discovering is performed without using any
knowledge about properties or features of sequences in the set of unaligned sequences.
- 20 5. The method of claim 1, further comprising the steps of:
if the candidate sequence comprises the predetermined number of patterns, adding the
candidate sequence to the set of sequences to create a new set of sequences; and
performing the step of discovering on the new set of sequences.
- 25 6. The method of claim 1, wherein each sequence comprises a series of symbols and
wherein each pattern comprises a plurality of positions, some of the plurality of positions each comprise
at least one expected symbol and other of the plurality of positions comprise positions which may be
occupied by any sequence character.
- 30 7. The method of claim 6, wherein, for one of the positions, the at least one expected
symbol is a plurality of expected symbols.

8. The method of claim 3, wherein the step of determining if each of the plurality of patterns is statistically significant comprises the steps of selecting one of the patterns, determining if a probability that the selected pattern occurs in a sequence meets a predetermined threshold, and continuing to select additional patterns until each pattern has been selected.

5

9. The method of claim 8, wherein the step of determining if a probability that the selected pattern occurs in a sequence meets a predetermined threshold further comprises the steps of using a second-order Markov chain method to determine the probability that the selected pattern occurs in a sequence and determining a natural logarithm of the probability that the selected pattern occurs in a sequence.

10

10. The method of claim 3, wherein the step of determining if each of the plurality of patterns is statistically significant further comprises the steps of removing instances of each of the patterns from the set of sequences to create a new set of sequences and performing the step of discovering on the new set of sequences.

15

11. The method of claim 3, wherein the step of determining if each of the plurality of patterns is statistically significant further comprises the steps of if any of the patterns is statistically significant, selecting a statistically significant pattern, modifying a composite descriptor to include the selected pattern if the selected pattern is not already part of the composite descriptor, and continuing to select statistically significant patterns until all statistically significant patterns have been selected.

20

12. The method of claim 1, wherein the step of discovering a plurality of patterns common to a plurality of the sequences comprises the steps of:

25

selecting a predetermined threshold that indicates how many of the sequences should contain a pattern for the pattern to be considered common;

discovering patterns, if any, that are common to the predetermined threshold of sequences;

30

if there are no patterns common to the predetermined threshold of sequences, decreasing the predetermined threshold; and

performing, until the predetermined threshold is less than a predetermined amount, the step of discovering patterns, if any, that are common to the predetermined threshold of sequences and

the step of if there are no patterns common to the predetermined threshold of sequences, decreasing the predetermined threshold.

13. (Canceled)
14. (Canceled)
15. (Canceled)
16. (Canceled)
17. (Canceled)
18. (Canceled)
19. (Canceled)
20. (Canceled)
21. (Canceled)
22. (Canceled)
23. A system comprising:
a memory that stores computer-readable code; and
a processor operatively coupled to said memory, said processor configured to implement said computer-readable code, said computer-readable code configured to:
provide a set of sequences, wherein the sequences are not aligned;
discover a plurality of patterns common to a plurality of the sequences; and
determine if a candidate sequence comprises a predetermined number of the patterns.
24. (Canceled)

25. An article of manufacture comprising:
a computer readable medium having computer readable code means embodied thereon,
said computer readable program code means comprising:

5 a step to provide a set of sequences, wherein the sequences are not aligned;
a step to discover a plurality of patterns common to a plurality of the sequences; and
a step to determine if a candidate sequence comprises a predetermined number of the
patterns.

10 26. (Canceled)